Alvarez et al. Journal of Biomedical Semantics 2011, 2:3
http://www.jbiomedsem.com/content/2/1/3

JOURNAL OF BIOMEDICAL SEMANTICS

RESEARCH Open Access

A shortest-path graph kernel for estimating gene product semantic similarity

Marco A Alvarez[1], Xiaojun Qi[1] and Changhui Yan[2*]

* Correspondence: changhui.yan@ndsu.edu
[2]Department of Computer Science, North Dakota State University, Fargo, 58108, USA
Full list of author information is available at the end of the article

Abstract

Background: Existing methods for calculating semantic similarity between gene products using the Gene Ontology (GO) often rely on external resources, which are not part of the ontology. Consequently, changes in these external resources like biased term distribution caused by shifting of hot research topics, will affect the calculation of semantic similarity. One way to avoid this problem is to use semantic methods that are "intrinsic" to the ontology, i.e. independent of external knowledge.

Results: We present a shortest-path graph kernel (spgk) method that relies exclusively on the GO and its structure. In spgk, a gene product is represented by an induced subgraph of the GO, which consists of all the GO terms annotating it. Then a shortest-path graph kernel is used to compute the similarity between two graphs. In a comprehensive evaluation using a benchmark dataset, spgk compares favorably with other methods that depend on external resources. Compared with simUI, a method that is also intrinsic to GO, spgk achieves slightly better results on the benchmark dataset. Statistical tests show that the improvement is significant when the resolution and EC similarity correlation coefficient are used to measure the performance, but is insignificant when the Pfam similarity correlation coefficient is used.

Conclusions: Spgk uses a graph kernel method in polynomial time to exploit the structure of the GO to calculate semantic similarity between gene products. It provides an alternative to both methods that use external resources and "intrinsic" methods with comparable performance.

Background

The Gene Ontology (GO) [1] systematically organizes knowledge by means of well-structured controlled vocabularies and provides consistent descriptions to organisms across species. GO terms have been widely used to annotate genes and gene products in the Gene Ontology Annotation (GOA) project [2]. As the GO becomes more and more important in biomedical research, computational methods are often needed to explore the GO to calculate the semantic similarity between gene products. Such methods have been used in a broad range of applications, including: clustering of genes in pathways [3-6], prediction of protein-protein interactions [7], and the evaluation of similarity between gene products with respect to expression profiles [8], protein sequence [9-11], protein function [12], and protein family [13].

The semantic similarity between two gene products is usually calculated based on the term similarity. First, pairwise semantic similarities between GO terms that annotate

the gene products are calculated. Then, the these pairwise similarities are combined to derive an overall semantic similarity between the gene products. Different methods have been used to combine pairwise GO term similarities in previous research [4,8,10,11,14,15]. A representative collection of methods for calculating the semantic similarity between GO terms has been reviewed in [16]. Most of those methods use the information content (IC) of the nearest common ancestor (NCA) or most informative common ancestor (MICA) to quantify the amount of shared information between two GO terms. However, the IC is calculated based on the frequency of GO terms in external resources, such as GOA databases. External resources change as knowledge is updated (e.g., more annotations are included in GOA). Consequently, for the same pair of GO terms, their semantic similarity computed by these methods might change as the external resources evolve. However, semantic similarities between GO terms should not be affected by such changes. In addition, certain annotations might be frequent simply because of popular research topics, leading to biased results. Some other methods rely on distance measures [17,18], e.g. counting the number of edges on the shortest path between the involved terms in the GO, to compute the GO term similarity. One shortcoming of this approach is that the edges in the GO do not imply equal length in semantics. Although some methods tried to address this issue by assigning different weights to edges at different levels, they still suffer from the fact that GO terms at the same level do not necessarily have the same specificity. Other methods calculate the semantic similarity between gene products without considering the semantic similarity between GO terms. In these methods, a gene product is represented by a set or a vector of GO terms that annotate it. Then, the semantic similarity between gene products is calculated as the overlap between sets or the inner product of vectors [4,10]. However, these methods did not exploit the structure of the GO and ignored the relationship between GO terms.

To address the aforementioned issues, we propose a shortest-path graph kernel (spgk) method for calculating the semantic similarity between gene products. In spgk, each gene product is represented as a graph, which is an induced subgraph of the GO. Then a graph kernel method is used to calculate the semantic similarity between the graphs. Spgk is intrinsic to the GO, i.e., it does not rely on external resources to calculate the semantic similarity. Thus, it does not have the same drawbacks as the methods based on the IC of GO terms. At the same time, it uses a graph to explicitly explore the GO structure and exploit the relationship between GO terms. Graph matching is computationally expensive in general, being an NP-complete problem on general graphs. To reduce the computational complexity, we develop a graph kernel to calculate the similarity between graphs. Using a comprehensive evaluation benchmark developed by another group, we compare spgk with other state-of-the-art methods.

Methods

In this section, we present a novel method for calculating the semantic similarity between proteins. First, we introduce basic background of the Gene Ontology. Then we describe the details of the graph kernel method.

Gene ontology and gene ontology annotations

The GO project [1] maintains a dynamic, structured, precisely defined, and controlled vocabulary of terms for describing the properties of gene products across species. The GO consists of three different ontologies describing: 1) biological processes (BP), where a process often involves a chemical or physical transformation (e.g. cell growth); 2) molecular functions (MF), where functions are defined as the biochemical activity of gene products (e.g. enzymes); and 3) cellular components (CC), which refers to places in the cell where gene products are active (e.g. nuclear membrane). Each ontology is structured as a directed acyclic graph, where nodes (GO terms) are linked to each other through "is-a", "part-of" or "regulates" relationships. On the other hand, the annotation of gene products is the process of assigning ontology terms to gene products in order to describe their activities and localization. For example, the GOA project [2], at the European Bioinformatics Institute (EBI), aims to provide high-quality electronic and manual annotations to UniProt KnowledgeBase (UniProtKB) entries [19]. GOA annotations are obtained from strictly controlled methods, where every association is supported by a distinct evidence source. A protein can be annotated with multiple GO terms from any of the three ontologies in the GO. Functional annotations of UniProtKB proteins currently consist of over 32 million annotations, which cover more than 4 million proteins [2].

Graph representation of proteins

We represented a protein using a subgraph of the ontology that consisted of all the GO terms annotating the protein and their ancestors in the ontology. Each edge of the graph corresponds to a relationship between two terms in the ontology. There are three types of relations in the GO: is-a, part-of, and regulates. Since the GO includes three different ontologies, the resulting graph will be different when a different ontology is used. For example, Figure 1 shows the graph generated for UniprotKB protein P17252, using the Cellular Component (CC) ontology.

A shortest-path graph kernel for proteins

We used a shortest-path graph kernel to compare two graphs as proposed in [20]. First, let's define the shortest-path graph. Given a graph $G = (V, E)$, its shortest-path graph is $G_{sp} = (V, E')$, where $E' = \{e'_1,...,e'_l\}$ such that $e'_i = (u, v)$, where $u \in V$, $v \in V$, and $path(u, v) \neq 0$. That is, G_{sp} has the same vertices as G and the edge (u, v) in G_{sp} has the same length as the shortest distance between u and v in G. This transformation can be performed using any all-pairs shortest path algorithm. In particular, the Floyd-Warshall algorithm is used in spgk because it is straightforward and has time complexity of $O(n^3)$. Then, for a pair of graphs, the shortest-path kernel calculates their similarity by comparing every pair of edges in their shortest-path graphs. For example, Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs and $G_{1sp} = (V_1, E'_1)$ and $G_{2sp} = (V_2, E'_2)$ be their shortest-path graphs respectively. The similarity between G_1 and G_2 can be calculated using Eq. 1.

$$K_{sp}(G_1, G_2) = \sum_{e_1 \in E'_1} \sum_{e_2 \in E'_2} (k_{walk}(e_1, e_2)) \quad (1)$$

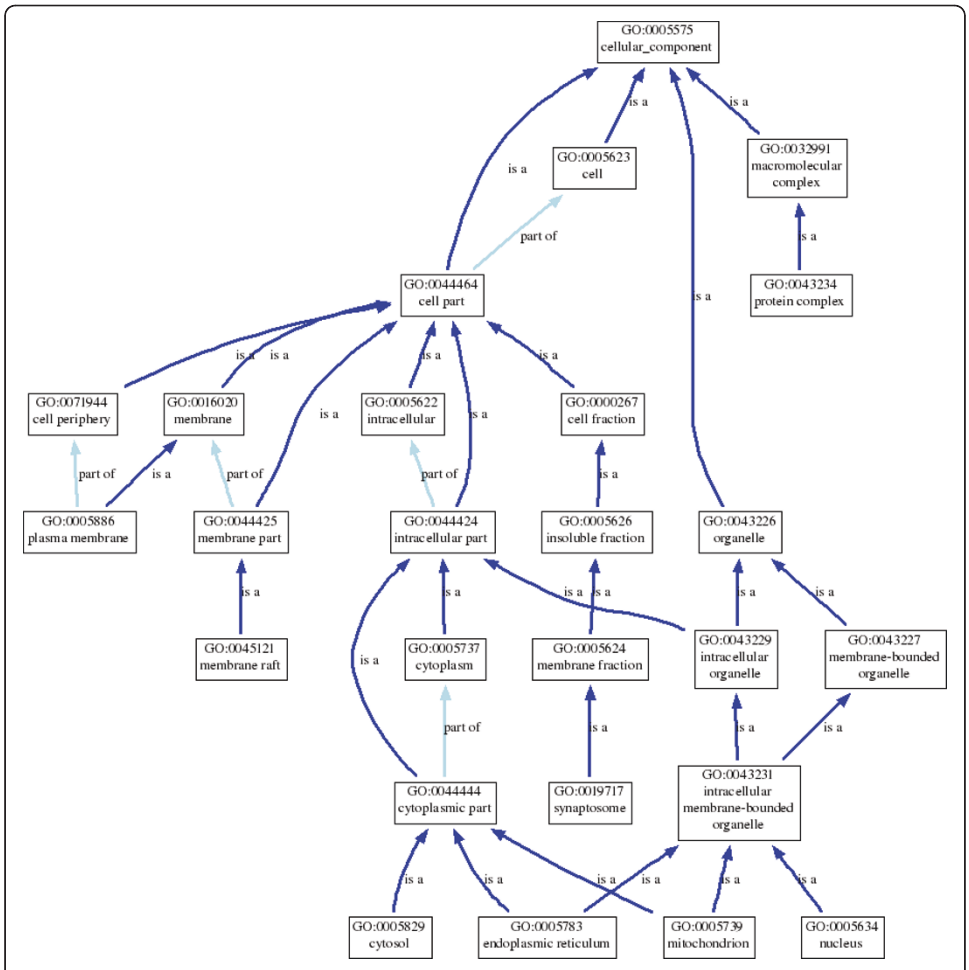

Figure 1 A subgraph generated from GO. The subgraph consisting of terms annotating protein P17252 (Protein kinase C alpha type) and their ancestors in the Cellular Component ontology.

where k_{walk} is a positive definite kernel for comparing two walks. In this case, a walk includes an edge and its two end nodes. Let e_1 be the edge connecting nodes v_1 and w_1, and e_2 be the edge connecting nodes v_2 and w_2, then $k_{walk}(e_1, e_2)$ is defined by Eq. 2.

$$k_{walk}(e_1, e_2) = k_{node}(v_1, v_2) * k_{edge}(e_1, e_2) * k_{node}(w_1, w_2) \quad (2)$$

where k_{node} is a kernel function for comparing two nodes, which returns 1 when the two nodes are identical and 0 otherwise, and k_{edge} is a kernel function for comparing two edges. k_{edge} is a Brownian bridge kernel that returns the largest value when two edges have identical length, and 0 when the edges differ in length more than a constant c as shown in Eq. 3. In this study, we use $c = 2$ as suggested by [20].

$$k_{edge}(e, f) = max(0, c - |length(e) - length(f)|) \quad (3)$$

Evaluation approach

We evaluated the performance of spgk by comparing the resulting semantic similarities with protein functional similarities derived from expert annotations. Functional

similarities between proteins were derived from the Pfam database [21] as described by Couto et al. [13]. Let P denote a protein and $F(P) = \{f_1, f_2, ..., f_n\}$ be the set of Pfam families that P is associated with. Then the functional similarity between two proteins P_i and P_j is given by Eq. 4

$$FS_f(P_i, P_j) = \frac{|F(P_i) \cap F(P_j)|}{|F(P_i) \cup F(P_j)|} \quad (4)$$

Previous study by Xu et al. [7] shows that having more annotations per protein in the dataset leads to more reliable functional similarity estimation from the GO. Thus, for the purpose of evaluation, we carefully selected a set of 100 proteins from GOA, such that they were the top 100 proteins with the highest numbers of annotations. We also ensured that for any selected protein: 1) it existed in the UniProtKB/Swiss-Prot database, 2) it had at least one annotation from each of the three ontologies in GOA-Uniprot, and 3) it had at least one Pfam-A annotation. The evaluation proceeded as follows: First, the graph kernel was used to calculate pairwise semantic similarities for a set of proteins. Second, pairwise functional similarities between the proteins were calculated based on the Pfam database annotations. Last, the Pearson's Correlation Coefficient between the semantic and functional similarities was calculated. If two proteins have similar function, then a good semantic similarity method should detect high semantic similarity between them. Thus, higher values of Pearson's Correlation Coefficient indicate better performance in the calculation of the semantic similarity. This procedure was repeated for each of the three ontologies in the GO, namely, BP, MF, and CC.

Results and discussion

Datasets

In our experiments, we used the revision 1.723 of the GO and the release 74.0 of GOA-Uniprot, where GO terms are assigned to proteins in UniProtKB by manual and electronic methods [2]. As mentioned before, the GO contains three different ontologies that describe gene products in terms of their associated biological processes, molecular functions, and cellular components.

Performance of spgk

100 proteins with the most GOA annotations were selected as described in the Methods section. Spgk was used to calculate pairwise semantic similarities between the proteins. The correlation coefficient between the resulting semantic and functional similarities was calculated. The evaluation was repeated using three different ontologies of the GO. The results are shown in Table 1 which reveals a couple of interesting points. First, spgk produces semantic similarities that are highly correlated with functional similarities for all three ontologies. Second, when the CC ontology is used, the correlation coefficients are lower than when the MF and BP ontologies are used. This is not surprising because the MF and BP ontologies are directly related to functions while the CC ontology is related to cellular components and locations.

Comparison of spgk with state-of-the-art methods

To compare spgk with other existing methods, we used the Collaborative Evaluation of GO-Based Semantic Similarity Measures (CESSM) online tool [22]. This tool has been

Table 1 Performance of spgk.

Ontology	BP	MF	CC
Pearson's Correlation Coefficient	0.855	0.852	0.703

The performance is measured by the Pearson's correlation coefficients between the semantic similarity given by spgk and the functional similarity estimated from Pfam annotations. BP, MF, CC are the three ontologies in the GO.

made available by the XLDB research group at the University of Lisbon. For the purpose of comparisons, CESSM provides a standard dataset consisting of 13,340 pairs of proteins involving 1,039 distinct proteins and implements 11 state-of-the-art semantic similarity methods, namely, simGIC and simUI [9], and three versions (the average, maximum and best-match average) of three different term similarity methods, namely Resnik [23], Lin [24], and Jiang & Conrath [25]. As a result, users can compare their methods with the 11 methods using the standard dataset.

As pointed out by Pesquita et al. [9] in a comprehensive evaluation, the maximum and average versions of term similarity methods have limitations from a biological point of view. Comparisons using the standard datasets at CESSM also confirmed that the best-match average version has better performance than the maximum and average versions for Resnik [23], Lin [24] and Jiang & Conrath [25] methods. Thus, in this section, we will compare spgk with simGIC, simUI, and the best-match average version of Resnik [23], Lin [24] and Jiang & Conrath [25] methods using CESSM. CESSM provides three different ways for evaluating a semantic similarity method, i.e., comparing the resulting semantic similarities with (1) functional similarities measured as sequence similarities, (2) functional similarities derived from enzyme commission (EC) classification, and (3) functional similarities derived from Pfam annotations.

Since the MF ontology is more closely related to function than the BP and CC ontologies, we will use the MF ontology to compare different methods. As pointed out by Pesquita et al. [9], the relationship between the semantic similarity and the sequence similarity is not linear. Thus, they recommended to use resolution instead of correlation coefficient to evaluate how well the semantic similarity matches the sequence similarity. Based on their definition, resolution is the relative intensity where variations in the sequence similarity scale are translated into the semantic similarity scale. Higher resolution values mean that the semantic similarity method has a higher capability to distinguish between different levels of protein functions. Therefore, a method with a higher resolution performs better than a method with a lower resolution. Table 2 shows the resolutions for different methods when the sequence similarity is compared with the semantic similarity computed by the methods. When the semantic similarity is compared with the function similarity derived from the EC classification and Pfam annotations, the Pearson's correlation coefficient is used as described in Methods. Tables 3 and 4 show the results.

The spgk method achieves the best results in tables 2 and 3, and is the second best in table 4. In addition to the better performance, the key advantage of spgk is that it is intrinsic to the ontology, i.e., it does not rely on external resources in the calculation of the semantic similarity. In contrast, all the other methods (except simUI) shown in tables 2, 3 and 4, rely on external resources, i.e., the annotations in GOA. Despite the high computational cost associated with the general graph comparisons, spgk does not suffer from this drawback. Using the shortest-path graph kernel, spgk requires a polynomial time ($O(n^4)$), where n is the number of vertices. In additioin, each step of the

Table 2 Comparison I.

Method	Resolution
spgk	0.976
simUI	0.967
Resnik	0.958
simGIC	0.956
Lin	0.571
Jiang & Conrath	0.241

The performance is measured by the resolution score.

graph kernel is simple to compute. For example, k_{node} only needs to compare whether two vertex IDs are identical, and k_{edge} considers the length difference between two edges. Thus, the constant factors associated with the polynomial time complexity are very small and spgk can run very fast in real applications.

SimUI is also intrinsic to the ontology. In simUI, the semantic similarity between two proteins is defined as the fraction between the number of GO terms shared by the two proteins and the number of GO terms in their union. Thus, simUI requires only a linear time ($O(n)$) and has the advantage that it is simple and faster for calculation. However, tables 2, 3, 4 show that spgk slightly outperformed simUI in all cases. We estimated the statistical significance of the improvement of spgk over simUI using Fisher's transformation. The p values were less than 0.001 when resolution was used to measure performance (table 2), 0.0384 for the EC similarity correlation coefficient (table 3) and 0.2266 for the Pfam similarity correlation coefficient (table 4). Therefore, compared with the conventional threshold of 0.05, the improvement is significant when the performance is measured by resolution and EC similarity correlation coefficient, but is insignificant when measured by Pfam similarity correlation coefficient. Comparing tables 2, 3, 4, we can see that the performance in table 4 is the poorest for all the methods. That might partially explain why the improvement is insignificant when Pfam similarity correlation coefficient is used as the measurement (table 4).

Conclusions

In this manuscript, we have presented a method (spgk) that computes the semantic similarity between gene products using only information intrinsic to GO. In comprehensive evaluations using a benchmark dataset, spgk compares favorably with other state-of-the-art methods that depend on external resources. Compared to simUI, spgk achieves slightly better results but also has a higher time complexity. A big difference between spgk and simUI is that spgk takes into account the structure of the ontology. Since the structure of the ontology contains important information, it is important to

Table 3 Comparison II.

Method	EC Similarity
spgk	0.646
Lin	0.642
simUI	0.637
simGIC	0.622
Resnik	0.603
Jiang & Conrath	0.561

The performance is measured by the Pearson's correlation coefficient between the resulting semantic similarity and the functional similarity derived from the EC classification.

Table 4 Comparison III.

Method	Pfam Similarity
simGIC	0.638
spgk	0.622
simUI	0.618
Resnik	0.572
Lin	0.564
Jiang & Conrath	0.491

The performance is measured by the Pearson's correlation coefficient between the resulting semantic similarity and the functional similarity derived from Pfam annotations.

exploit them to capture semantic similarity. The results presented here show that spgk provides an alternative to both methods that rely on external resources and "intrinsic" methods with comparable performance.

In light of future development, there are still some limitations in spgk at its current form. For example, in spgk, the function (k_{node}) that compares nodes only considers whether the two nodes are identical. However, each node in the GO is associated with a text definition, which contains rich information that is useful for deriving biological relationship between nodes. Thus, one direction for future improvement is to take into account the semantics of the text definition when comparing nodes. Furthermore, the k_{edge} function only considers the length difference between two paths. In GO, the edges are associated with different types of relationship. Since different types of relationship have different biological meanings, they should be given different weights. Thus, another direction for improvement is to systematically explore weighting methods that assign different weights to the edges based on the biological relationships.

Acknowledgements

We would like to thank the XLDB Research Team from the University of Lisbon for providing an online tool for the evaluation of GO-based semantic similarity measures. In particular, we thank Catia Pesquita for all the kind support given for using their tool. This project was partially supported by NIH Grant Number P20 RR016471 from the INBRE Program of the National Center for Research Resources.

Author details

[1]Department of Computer Science, Utah State University, Logan, 84322, USA. [2]Department of Computer Science, North Dakota State University, Fargo, 58108, USA.

Authors' contributions

CY conceived the project and supervised all aspects of the research. MA contributed to programming, discussion, data analysis and preparation of the first draft. XQ contributed to discussion. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 27 February 2011 Accepted: 29 July 2011 Published: 29 July 2011

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
2. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009–an integrated Gene Ontology Annotation resource.** *Nucl Acids Res* 2009, **37**:D396-403.
3. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F: **A new method to measure the semantic similarity of go terms.** *Bioinformatics* 2007, **23**:1274-1281.
4. Sheehan B, Quigley A, Gaudin B, Dobson S: **A relation based measure of semantic similarity for gene ontology annotations.** *BMC Bioinformatics* 2008, **9**:468.
5. Nagar A, Al-Mubaid H: **A new path length measure based on go for gene similarity with evaluation using sgd pathways.** *Proceedings of IEEE International Symposium on Computer-Based Medical Systems* 2008, 590-595.

6. Du Z, Li L, Chen C-F, Yu PS, Wang JZ: **G-sesame: web tools for go-term-based gene similarity analysis and knowledge discovery.** *Nucl Acids Res* 2009, **37**:W345-349.
7. Xu T, Du L, Zhou Y: **Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data.** *BMC Bioinformatics* 2008, **9**:472.
8. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A: **Correlation between gene expression and go semantic similarity.** *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2005, **2**:330-338.
9. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcão AO, Couto FM: **Metrics for go based protein semantic similarity: a systematic evaluation.** *BMC Bioinformatics* 2008, **9**:5.
10. Mistry M, Pavlidis P: **Gene ontology term overlap as a measure of gene functional similarity.** *BMC Bioinformatics* 2008, **9**:327.
11. Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19**:1275-1283.
12. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S: **Rapid Annotation of Anonymous Sequences from Genome Projects Using Semantic Similarities and a Weighting Scheme in Gene Ontology.** *PLoS ONE* 2009, **4**:e4619.
13. Couto FM, Silva MJ, Coutinho PM: **Measuring semantic similarity between gene ontology terms.** *Data and Knowledge Engineering* 2007, **16**:137-152.
14. Schlicker A, Domingues F, Rahnenfuhrer J, Lengauer T: **A new measure for functional similarity of gene products based on Gene Ontology.** *BMC Bioinformatics* 2006, **7**:302.
15. Alvarez M, Qi X, Yan C: **GO-Based Term Semantic Similarity.** In *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*. Edited by: Wong W, Liu W, Bennamoun M. Pennsylvania: IGI-Global; 2011:174-185.
16. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM: **Semantic similarity in biomedical ontologies.** *PLOS Computational Biology* 2009, **5**:e1000443.
17. Cheng J, Cline M, Martin J, Finkelstein D, Awad T, Kulp D, Siani-Rose MA: **A knowledge-based clustering algorithm driven by gene ontology.** *J Biopharm Stat* 2004, **14**:687-700.
18. Wu X, Zhu L, Guo J, Zhang D-Y, Lin K: **Prediction of yeast proteinprotein interaction network: insights from the gene ontology and annotations.** *Nucl Acids Res* 2006, **34**:2137-2150.
19. The UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010.** *Nucl Acids Res* 2010, **38**:D142-148.
20. Borgwardt KM, Ong CS, Schonauer S, Vishwanathan SVN, Smola AJ, Kriegel H-P: **Protein function prediction via graph kernels.** *Bioinformatics* 2005, **21**:i47-56.
21. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, H-R Hotz, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The pfam protein families database.** *Nucl Acids Res* 2008, **36**:D281-288.
22. Pesquita C, Pessoa D, Faria D, Couto F: **CESSM: Collaborative Evaluation of Semantic Similarity Measures.** *Proceedings of JB2009: Challenges in Bioinformatics Lisbon, Portugal* 2009.
23. Resnik P: **Using information content to evaluate semantic similarity in a taxonomy.** *Proceedings of International Joint Conference on Artificial Intelligent* 1995, 448-453.
24. Lin D: **An information-theoretic definition of similarity.** *Proceedings of International Conference on Machine Learning* 1998, 296-304.
25. Jiang JJ, Conrath DW: **Semantic similarity based on corpus statistics and lexical taxonomy.** *Proceedings of International Conference Research on Computational Linguistics* 1997, 19-33.

doi:10.1186/2041-1480-2-3
Cite this article as: Alvarez *et al.*: A shortest-path graph kernel for estimating gene product semantic similarity. *Journal of Biomedical Semantics* 2011 **2**:3.